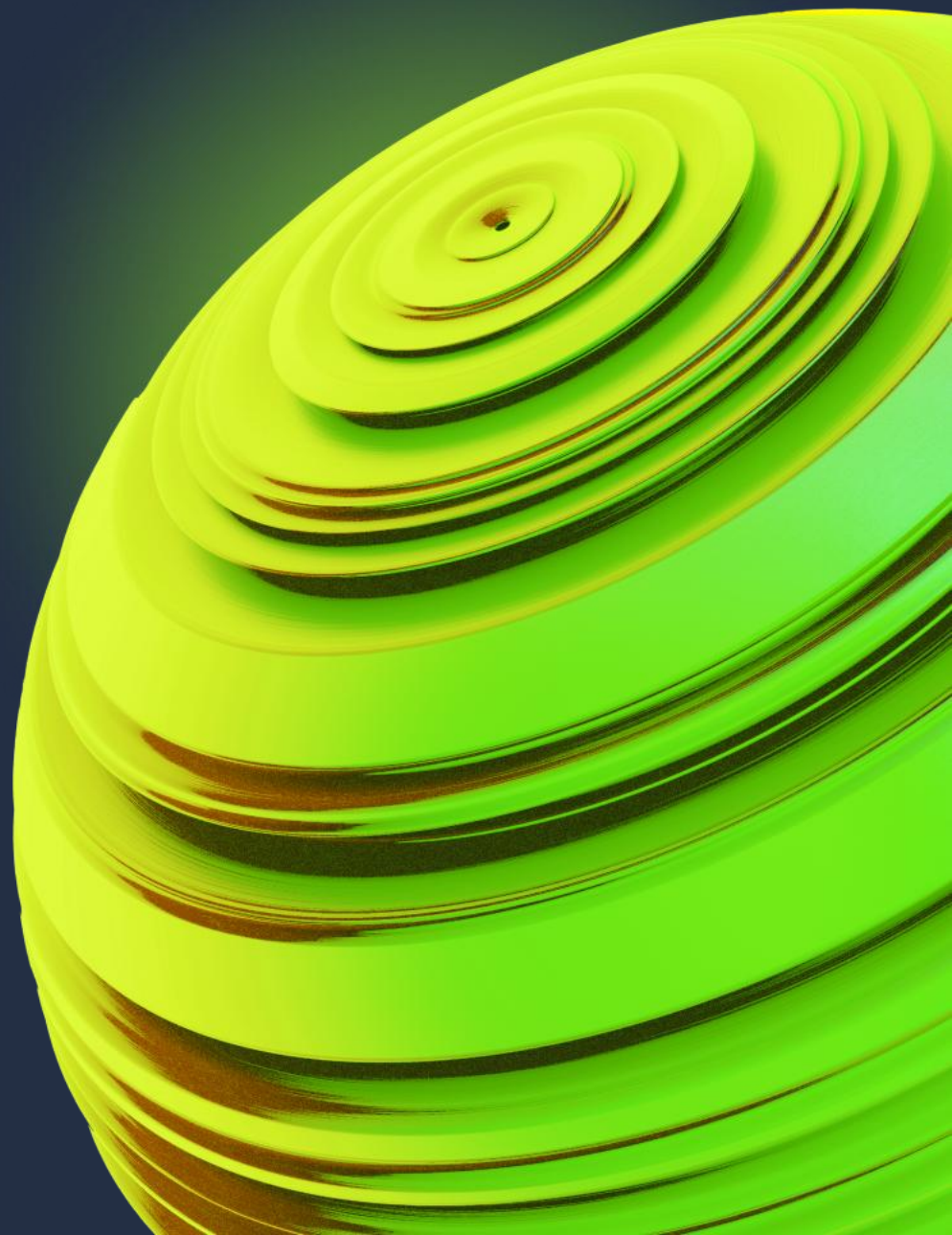
ИНСТИТУТ ДОПОЛНИТЕЛЬНОГО ОБРАЗОВАНИЯ
УНИВЕРСИТЕТА ИННОПОЛИС

УНИВЕРСИТЕТ ИННОПОЛИС

Итоговый проект: Прогнозирование цен на недвижимость с использованием методов машинного обучения

Работу выполнила Татаринова Анастасия Павловна

Дата: сентябрь 2025



Цели и задачи

Цель проекта:

Создать модель, способную спрогнозировать цены на квартиры в Москве на основе имеющихся данных.

Задачи:

- Проанализировать доступный датасет с характеристиками квартир.
- Выявить значимые признаки, влияющие на цену.
- Обучить и сравнить несколько моделей машинного обучения.
- Реализовать удобный способ использования модели для прогнозирования цены.
- Сформулировать рекомендации для дальнейшего развития.

Сравнение сервисов прогнозирования цен на недвижимость

Сервис	Описание	Количество параметров	Особенности	Доступность
Домклик (Сбербанк)	Онлайн оценка с учётом ремонта, типа дома	Более 10	Бесплатный, прогноз времени продажи	Веб, мобильное приложение
IRN.ru	Индекс и прогнозы рынка Москвы, Подмосковья	Среднее	Использует макроэкономические данные	Веб-сайт
ЦИАН	Точный расчет с детализацией по ремонту, этажам	Большое	Учитывает год постройки и состояние	Веб-платформа, API
Авито Оценка	На основе предложений рынка и объявлений	Небольшое	Быстрая оценка по аналогам	Веб, мобильное приложение
Technologika	ИИ-прогнозы изменения стоимости на годы	Много	Высокая точность для корпоративных клиентов	Корпоративное решение
Tiqum	Анализ и прогнозы рынка для агентов	Разнообразное	Отчёты и данные для профессионалов	SaaS платформа

Выводы:

- Домклик и ЦИАН предоставляют подробные и бесплатные оценки для конечных пользователей.
- IRN.ru больше ориентирован на макро аналитику и прогнозы рынка в целом.
- Корпоративные и SaaS-сервисы (Technologika, Tiqum) предлагают расширенные аналитические возможности, но доступны не всем.

Датасет

Для реализации выбран датасет [Moscow Housing Price Dataset с Kaggle.]

- Основные признаки: тип квартиры, площадь, количество комнат, этаж, близость к метро, тип ремонта, регион, этаж, количество этажей в доме, ремонт и другие.
- Проведена очистка данных: удалены пропуски и дубликаты, проведена стандартизация и кодирование категориальных признаков.

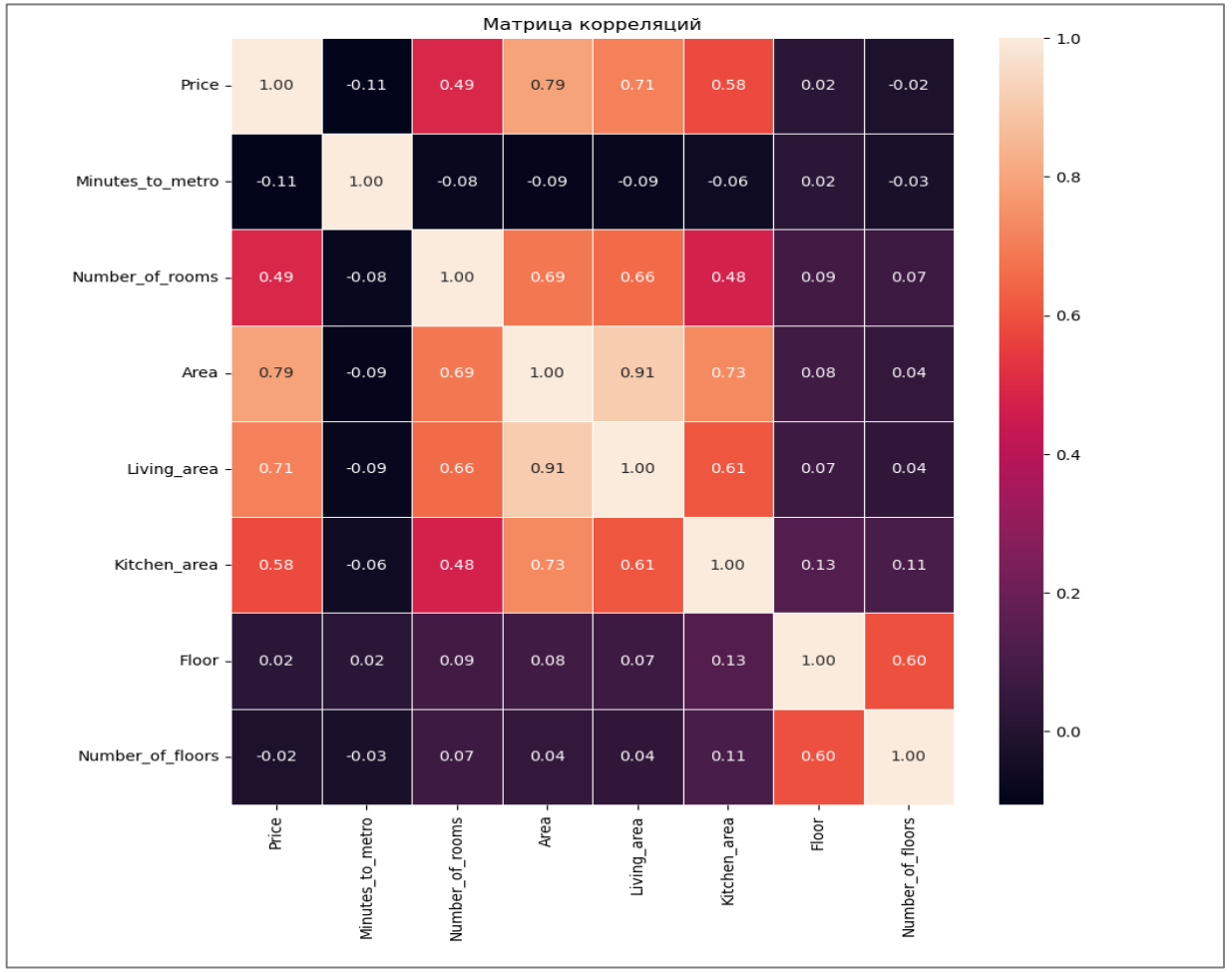
	Price	Apartment type	Metro station	Minutes to metro	Region	Number of rooms	Area	Living area	Kitchen area	Floor	Number of floors	Renovation
0	6300000.0	Secondary	Опалиха	6.0	Moscow region	1.0	30.6	11.1	8.5	25.0	25	Cosmetic
1	9000000.0	Secondary	Павшино	2.0	Moscow region	1.0	49.2	20.0	10.0	6.0	15	European-style renovation
2	11090000.0	Secondary	Мякинино	14.0	Moscow region	1.0	44.7	16.2	13.1	10.0	25	Cosmetic
3	8300000.0	Secondary	Строгино	8.0	Moscow region	1.0	35.1	16.0	11.0	12.0	33	European-style renovation
4	6450000.0	Secondary	Опалиха	6.0	Moscow region	1.0	37.7	15.2	4.0	5.0	5	Without renovation

	Minutes_to_metro	Number_of_rooms	Area	Living_area	Kitchen_area	Floor	Number_of_floors	Apartment_type_New building	Apartment_type_Secondary	Region_Moscc
0	6.0	1.0	30.60	11.1	8.5	25.0	25	0	1	
1	2.0	1.0	49.20	20.0	10.0	6.0	15	0	1	
2	14.0	1.0	44.70	16.2	13.1	10.0	25	0	1	
3	8.0	1.0	35.10	16.0	11.0	12.0	33	0	1	
4	6.0	1.0	37.70	15.2	4.0	5.0	5	0	1	
...	...	...	...	...	...	...	...	...	...	...
22670	8.0	1.0	44.17	24.5	10.3	4.0	17	1	0	
22672	25.0	1.0	31.60	10.1	12.2	11.0	15	1	0	
22673	30.0	0.0	18.00	15.0	8.1	17.0	17	1	0	
22674	14.0	2.0	36.39	22.0	6.6	12.0	14	1	0	
22675	8.0	2.0	56.13	32.0	5.0	10.0	17	1	0	

20841 rows × 15 columns

Анализ данных — корреляции

- Наиболее сильная положительная корреляция с ценой у площади квартиры и жилой площади (~0.7-0.8).
- Количество комнат также оказывает заметное влияние (~0.5).
- Время до метро — слабо отрицательно коррелирует.
- Этажность и количество этажей незначительно влияют.



Гипотезы и их проверка

- Гипотеза 1: средние цены новостроек и вторичного фонда различаются.
- Гипотеза 2: средние цены квартир с одним и двумя комнатами отличаются.
- Проверка с помощью t-тестов показала статистическую значимость различий (p < 0.05).

Гипотеза 1: Влияние типа квартиры на цену

H0 (нулевая гипотеза): Средняя цена квартир "Вторичного рынка" и "Новостроек" не различается.

H1 (альтернативная гипотеза): Средняя цена квартир "Вторичного рынка" значительно отличается от средней цены "Новостроек".

Гипотеза 2: Влияние количества комнат на цену

H0: Средняя цена квартир с 1 комнатой равна средней цене квартир с 2 комнатами.

H1: Средняя цена квартир с 1 комнатой отличается от средней цены квартир с 2 комнатами.

[56]:
```
alpha = 0.05
# Гипотеза 1: Средняя цена квартир вторичного рынка и новостроек одинаков
# H0: μ_Secondary = μ_New building (средние равны)
# H1: μ_Secondary ≠ μ_New building (средние различаются)

secondary_prices = df[df['Apartment_type'] == 'Secondary']['Price'].dropna()
new_building_prices = df[df['Apartment_type'] == 'New building']['Price'].dropna()

# Проверка: t-тест Стьюдента (две независимые выборки, нормальное распределение приближённо допустимо при n > 30)
t_stat, p_value = ttest_ind(secondary_prices, new_building_prices, equal_var=False)

print(f"Средняя цена квартир вторичного рынка: {np.round(secondary_prices.mean(), 2)}")
print(f"Средняя цена квартир новостроек: {np.round(new_building_prices.mean(), 2)}")
print(f"t-статистика: {np.round(t_stat, 3)}")
print(f"p-значение: {np.round(p_value, 5)}")
if p_value < alpha:
    print("Отвергаем нулевую гипотезу - средняя цена квартир вторичного рынка и новостроек различаются")
else:
    print("Не можем отвергнуть нулевую гипотезу - средняя цена квартир вторичного рынка и новостроек равны")
```

```
Средняя цена квартир вторичного рынка: 52255086.55
Средняя цена квартир новостроек: 8179494.21
t-статистика: 48.706
p-значение: 0.0
Отвергаем нулевую гипотезу - средняя цена квартир вторичного рынка и новостроек различаются
```

[58]:
```
alpha = 0.05
# Гипотеза 1: Средняя цена однокомнатных и двукомнатных квартир одинаков
# H0: μ_1 = μ_2 (средние равны)
# H1: μ_1 ≠ μ_2 (средние различаются)

one_prices = df[df['Number_of_rooms'] == 1]['Price'].dropna()
two_prices = df[df['Number_of_rooms'] == 2]['Price'].dropna()

# Проверка: t-тест Стьюдента (две независимые выборки, нормальное распределение приближённо допустимо при n > 30)
t_stat, p_value = ttest_ind(one_prices, two_prices, equal_var=False)

print(f"Средняя цена однокомнатных квартир: {np.round(one_prices.mean(), 2)}")
print(f"Средняя цена двукомнатных квартир: {np.round(two_prices.mean(), 2)}")
print(f"t-статистика: {np.round(t_stat, 3)}")
print(f"p-значение: {np.round(p_value, 5)}")
if p_value < alpha:
    print("Отвергаем нулевую гипотезу - средняя цена однокомнатных и двукомнатных квартир различаются")
else:
    print("Не можем отвергнуть нулевую гипотезу - средняя цена однокомнатных и двукомнатных квартир равны")
```

```
Средняя цена однокомнатных квартир: 10931651.78
Средняя цена двукомнатных квартир: 15015706.24
t-статистика: -14.612
p-значение: 0.0
Отвергаем нулевую гипотезу - средняя цена однокомнатных и двукомнатных квартир различаются
```

Обучение моделей

- Использовали модели: линейная регрессия, KNN, дерево решений, случайный лес и градиентный бустинг.
- Подобрали оптимальные гиперпараметры через Grid Search.
- Оценивали качество по метрикам R2, MSE и RMSE.

Результаты сравнения моделей

Модель	R2	RMSE
Линейная регрессия	0.64	44 583 582
KNN	0.67	42 817 634
Дерево решений	0.65	44 158 841
Случайный лес	0.74	37 699 576
Градиентный бустинг	0.72	39 219 532

Реализация решения

- Модель сохранена в файл с помощью библиотеки joblib.
- Реализован код в Jupyter для взаимодействия с пользователем через ввод признаков и выдачу прогноза.

```
        area,
        living_area,
        kitchen_area,
        floor,
        number_of_floors,
        apartment_type_new_building,
        apartment_type_secondary,
        region_moscow,
        region_moscow_region,
        renovation_cosmetic,
        renovation_designer,
        renovation_european,
        renovation_without
    ]])

    return user_data

if __name__ == "__main__":
    user_data = get_user_input()
    prediction = model.predict(user_data)
    print(f"Прогнозируемая цена недвижимости: {prediction[0]:,.2f}".replace(',', ' ') + " рублей")
```

```
Введите время до метро (в минутах):  3
Введите количество комнат:  2
Введите площадь квартиры (кв.м):  40
Введите жилую площадь (кв.м):  25
Введите площадь кухни (кв.м):  18
Введите этаж:  5
Введите количество этажей в доме:  10
Тип квартиры: 1 - New building (Новостройка), 0 - Secondary (Вторичка)
Введите 1 или 0:  1
Регион: 1 - Moscow, 0 - Moscow region
Введите 1 или 0:  0
Тип ремонта:
 1 - Cosmetic (Косметический)
 2 - Designer (Дизайнерский)
 3 - European-style (В европейском стиле)
 4 - Without renovation (Без ремонта)
Введите номер типа ремонта (0-4):  4
Прогнозируемая цена недвижимости: 8 470 449.69 рублей
```

Вывод

- Успешно разработана модель для прогноза цен по ключевым параметрам недвижимости.
- Выявлены значимые факторы, подтверждены гипотезы значимости.
- Случайный лес продемонстрировал более высокий уровень точности.
- Предусмотрена простая возможность использования модели для пользователей.

Дальнейшее развитие

- Расширить объем и качество признаков (например, добавить в обучение станцию метро).
- Использовать более сложные ансамбли и методы обучения.
- Автоматизация обновления модели на новых данных.
- Визуализация и расширение пользовательского интерфейса.

Литература

- Прогноз рынка недвижимости Москвы, Подмосковья и России до конца 2025 и на 2026 год от IRN.RU
- Бесплатная оценка стоимости недвижимости, рассчитать стоимость квартиры и жилья в калькуляторе онлайн – Домклик
- 5 сервисов, где можно бесплатно узнать стоимость квартиры
- Предсказание цен на недвижимость при помощи ИИ | Технологика
- Сервис анализа рынка недвижимости | Создание стартапов и цифровых продуктов для бизнеса: сайты, приложения и UX — Tiqum
- Топ-5 сервисов для оценки стоимости квартиры или объектов недвижимости
- Scikit-learn Documentation: https://scikit-learn.org
- Pandas Documentation: https://pandas.pydata.org
- NumPy Documentation: https://numpy.org
- Материалы курса «Искусственный интеллект и основы аналитики (больших) данных»

ИНСТИТУТ
ДОПОЛНИТЕЛЬНОГО
ОБРАЗОВАНИЯ
УНИВЕРСИТЕТА ИННОПОЛИС

Спасибо за внимание!

Контакты